Protein Engineering vol.8 no.4 pp.329–338, 1995

DNA recognition and superstructure formation by helix–turn–helix proteins

Masashi Suzuki[1,2], Naoto Yagi[3] and Mark Gerstein[4]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK, [3]Tohoku University, School of Medicine, Seiryo-machi, Sendai, 980-77 Japan and [4]Structural Biology Department, Fairchild Building, Stanford University, Stanford, CA 94305-5400, USA

[2]To whom correspondence should be addressed

The way helix–turn–helix proteins recognize DNA is analysed by comparing their sequences, structures, and binding specificities. Individual recognition helices in these proteins bind to four DNA base pairs with the same geometry. However, pairs of recognition helices in the protein dimers can have different separations and orientations. These differences are used for discriminating between DNAs which have different superstructures, in particular, different numbers of base pairs between sets of the four base pairs.

Key words: CAP/DNA bending/FIS/helix–turn–helix/NMR/transcription factor

Introduction

Helix–turn–helix (HTH) is an extensively characterized DNA-binding motif (Ohlendorf *et al.*, 1982; Sauer *et al.*, 1982; Pabo and Sauer, 1984; Pabo *et al.*, 1990; Brennen, 1991; Ptashne, 1992). However, a number of important questions still remain unanswered concerning its DNA recognition mode. In particular, there seems to be disagreement among scientists whether HTH proteins recognize DNA by the same binding geometry or not. Some scientists state that 'the mode of interaction with DNA can vary substantially' (Wilson *et al.*, 1992 ; see also Matthews, 1988) and show that overall DNA–HTH protein interactions look different from each other (see Figure 7.27 in Branden and Tooze, 1991; Figure 4 of Stetz, 1993), while others believe the same rules can explain the DNA-binding specificity of HTH proteins generally (Kisters-Woike *et al.*, 1991; Lehming *et al.*, 1991; Suzuki and Yagi, 1994). Also, if HTH proteins recognize DNA by the same binding geometry, the question arises why CAP bends DNA considerably (Schultz *et al.*, 1991), while other HTH proteins do not.

We address the above questions in terms of stereochemical principles which govern DNA recognition by HTH proteins. We show that individual recognition helices of classic HTH proteins bind to DNA in the same way but that pairs of HTH recognition helices can be combined in different ways in dimers and that these differences are used for discriminating between DNAs which have different superstructures. In particular we explain how CAP bends the DNA. In other words the disagreement was caused by not separating the two different levels of DNA recognition.

This analysis became possible by defining direct binding sites (here each binding site is defined as the four base pairs contacted by amino acid side chains) and recognition helices (here each helix is defined as the two helical turns which contact the DNA bases) appropriately and by describing the geometry of the two binding sites and that of the two recognition helices precisely.

Materials and methods

The structures of HTH proteins and those of DNA–HTH complexes, which have been determined by X-ray crystallography or NMR, are listed in Figure 1. The geometry of pairs of recognition helices relative to each other in a HTH dimer were characterized by four 'combination' parameters (Figure 2).

(i) *l*—line *l* is drawn connecting the centres of the two recognition helices and the length of this line is parameter *l*. The centre of each recognition helix is defined, as previously (M.Suzuki and M.Gerstein, submitted), as the projection onto the α-helix axis of the midpoint of a line connecting the C_α atoms at residue positions 3 and 4 (see the numbering in Figure 3).

(ii) δ—parameter δ measures the rotation around the recognition helix axis (Figure 2a). It is defined as 90°–θ, in which θ is the angle between line *l* and the line connecting the helix centre to the Cα atom of position 4 when looking down the helix axis. (Note, position 4 is occupied by a hydrophobic residue that points in towards the protein.)

(iii) ω—parameter ω measures the rotation of the two recognition helices relative to each other around line *l* (Figure 2b and c). 2ω is defined as the angle between the recognition helix axes when looking down *l*.

(iii) ε—parameter ε measures the tilt of the recognition helix relative to line *l* (Figure 2d). It is defined as the complement of the angle between the recognition helix axis and line *l*, projected onto a plane that contains line *l* and is inclined so that it bisects angle 2ω, i.e. so that it lies at an angle ω from each helix axis. More precisely, if the axes of the two recognition helices are $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ and vector **l** lies on line *l* and runs from helix 1 to helix 2, the normal to projection plane will be $\mathbf{l} \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)$.

We will make available electronically supplementary pictures relevant to our calculations (e-mail address: mbg@hyper.stanford.edu or URL: ftp://hyper.stanford.edu/pub/mbg/DNA/).

The parameters characterizing the DNA base pair steps—helical twist, roll, etc.—were calculated using a computer program developed by Babcock *et al.* (1993). The values calculated by this program conform to the Cambridge code of DNA parameters (Diekmann, 1989). The DNA in the CAP structure has two nicks near its centre and so it was not possible to use the program to calculate its base pair parameters for the two middlemost steps.

Results and discussion

DNA recognition by individual HTH recognition helices

Three types of residues are arranged into a HTH recognition helix. The residues in a HTH recognition helix have three different functional roles: (i) some residues contact DNA bases (these are important for the binding specificity), (ii) some residues contact DNA phosphates (these fix the binding geo-

HTH proteins

	Resolution	R factor	DNA	PDB code	reference
λR	3.2Å	0.22	---	---	Pabo and Lewis,1992
λR	1.8Å	0.19	17	1LMB	Clarke et al.,1991
λR	2.5Å	0.24	20	---	Jordan and Pabo,1988
λC	2.8Å	0.45	---	1CRO	Anderson et al.,1981
λC	3.9Å	0.50	17	4CRO	Brennan et al.,1990
434R	NMR	---	---	2PRA	Neri et al.,1992
434R	2.0Å	0.19	---	1R69	Mondragón et al.,1989b
434R	3.2Å	0.30	18	---	Anderson et al.,1987
434R	2.5Å	0.18	20	2OR1	Aggarwal et al.,1988
434R	2.5Å	0.19	20	1PER	Rodgers and Harrison,1993
434R	2.5Å	0.21	20	1RPE	Shimon and Harrison,1993
434C	2.4Å	0.20	---	2CRO	Mondragón et al.,1989a
434C	3.2Å	0.27	14	---	Wolberger et al.,1988
434C	2.5Å	0.22	20	3CRO	Mondragón and Harrison,1991
CAP	2.5Å	0.21	---	3GAP	Weber and Steitz,1987
CAP	3.0Å	0.24	31	1CGP	Schultz et al.,1991
BirA	2.3Å	0.19	---	1BIA	Wilson et al.,1992
BirA	2.3Å	0.19	---	1BIB	Wilson et al.,1992
P22R	NMR	---	---	1ADR	Sevilla-Sierra et al.,1994
Hin	2.3Å	0.23	14	1HCR	Feng et al.,1994
LacR	NMR	---	11	1LCC	Chuprina et al.,1993
TetR	2.5Å	0.20	---	---	Hinrichs et al.,1994
FIS	2.0Å	0.25	---	1FIA	Kostrewa et al.,1991
FIS	2.3Å	0.18	---	3FIS	Yuan et al.,1991
Oct1 POU					
	NMR	---	---	---	Assa-Munt et al.,1993
	NMR	---	---	---	Dekker et al.,1993
	3.0Å	0.24	15	1OCT	Klemm et al.,1994

HTH related

type 2					
TrpR [-trp]	1.8Å	0.20	---	3WRP	Zhang et al.,1987
TrpR [+trp]	2.2Å	0.20	---	1WRP	Schevitz et al.,1985
TrpR [+trp]	1.7Å	0.18	---	2WRP	Lawson et al.,1988
TrpR	1.9Å	0.17	19	1TRO	Otwinowski et al.,1988
TrpR	2.4Å	0.22	19	1TRR	Lawson and Carley,1993
type 3					
Engl	2.8Å	0.23	21	1HDD	Kissinger et al., 1990
Matα2	2.7Å	0.27	21	---	Wolberger et al.,1991
Antp	NMR	---	14	1AHD	Billeter et al.,1993
Oct1 homeo					
	3.0Å	0.24	15	1OCT	Klemm et al.,1994
type 4					
LexA	NMR	---	---	1LEA	Lamerichs et al.,1989
Myb	NMR	---	---	1MSE	Ogata et al., 1992
type 5					
HNF3	2.5Å	0.21	13	---	Clark et al., 1993
HSF	1.8Å	0.19	---	2HTS	Harrison et al.,1994

Fig. 1. Listing of crystal and NMR structures of HTH (top) and HTH-related (bottom) proteins. The number of base pairs in complexes with DNA is shown in the 'DNA' column. All coordinates were taken from the Protein Data Bank (Bernstein *et al.*, 1977).

metry) and (iii) the others face away from the DNA and thus can interact with the rest of the protein (these do not interact with the DNA but limit the rotation of the recognition helix).

We number the positions so that the hydrophobic residue (of type C) which is most important for packing against the preceding helix lies at position 4 (Figure 3). Positions 7 and 8, which are on the same phase as 4, are also often occupied by hydrophobic residues (i.e. type iii). In the crystal and NMR structures of the 434 cro protein (434C), the 434 repressor (434R), the λ repressor (λR), the Lac repressor (LacR), CAP and the Hin recombinase (Hin), residues at positions 1, 2 and 6 contact DNA bases and are classified as of (i). Position 5 also faces the DNA bases (i); however, the distance to the DNA bases from this position is larger and thus the contacts it makes are weaker and less frequent (see legend for Figure 4). Positions −1 (which is N-terminal to the recognition helix)

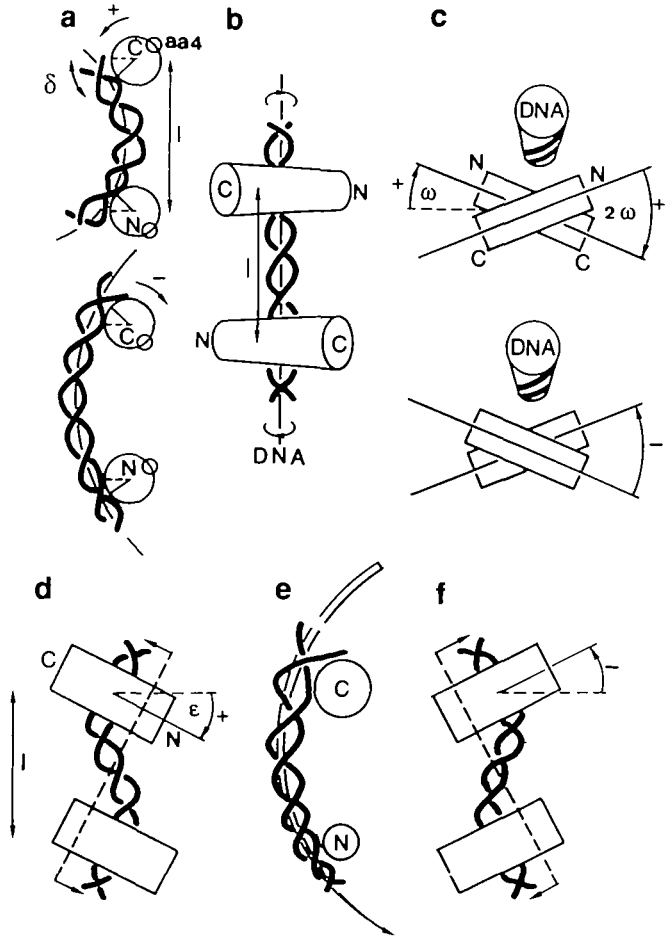


Fig. 2. Definition of 'combination' parameters for describing the relative orientation/separation of two recognition helices in a HTH dimer. The figure shows how these four parameters are related to features of the DNA superstructure: *l* to the spacing number, δ to the concave/convex curvature, ω to the twist and ε to the left/right-handedness of the DNA superhelicity (see text). (**a**) Parameter δ. The position of amino acid 4 is shown (aa4). (**b**) and (**c**) Parameter ω. In (c) one is looking down the DNA axis while in (b) one is looking perpendicular to it. (**d**)–(**f**) Parameter ε. When ε and δ have the same sign, these create a left-handed superhelix of DNA (e).

and 9 are often used for phosphate-contacting (i.e. type ii). Position 3 is sandwiched between a type (i) residue at position 2 and a type (iii) residue at position 4. It is very close to the sugar–phosphate backbone of the DNA and is occupied only by one of the small residues, Gly, Ala, Ser, Cys and Thr; position 3 is regarded as being of type (ii).

The above way of combining the three types of residue positions into a single helix is kept the same among the HTH proteins (Figure 3) and this produces the DNA-bindng geometry specific to the HTH proteins.

The Trp repressor (TrpR) binds to DNA in a different way. It is discussed separately later in this paper.

Contacts between amino acid side chains and DNA bases fall into the same pattern. For understanding the binding specificity, contacts between DNA bases and type (i) residues in the crystal and NMR structures were further analysed. These contacts can be summarized into 'charts'; each chart is a schematic sketch of the DNA major groove to which the recognition helix binds [Figure 4a–f and l; see Figure 5b for the Watson(W)–Crick(C) notation for the two DNA strands, the 5′–3′ direction of the strands and the N to C direction of the recognition helix in the centre]. We found that the contacts in the 434C, 434R, λR, LacR, CAP and Hin structures fall into a common pattern (Figure 5b).

By using the pattern of base–residue contacts deduced from the crystal and NMR structures it is now possible to examine the DNA-binding specificity of other HTH proteins, of which the DNA complex structures are as yet unknown (Figure 4); the structures of P22R, TetR and FIS have been determined only in the absence of DNA, λ cro (λC) has been crystallized in the presence of DNA but the resolution of the structure is not very high and Lehming *et al.* (1991) have listed DNA-binding specificity of many other HTH proteins of which the structures are unknown but their sequences are undoubtedly of the HTH fold. For this analysis understanding of two fundamental aspects of DNA–protein interactions are important.

First, the 20 amino acid residues can be classified into four groups, according to their shape and size (Suzuki, 1994): small, medium, large and aromatic. An amino acid side chain contacting a nucleotide base is, obviously, most easily replaced by one of similar size and shape. If, however, it is replaced by one of different size, it may contact nucleotide bases at different positions. Therefore, the pairings of residue and base positions in the stereochemical chart must be understood with specification of the size of the residue (Figure 5b).

Second, any contact between an amino acid side chain and the major groove side of a DNA base involves hydrogen bonding or hydrophobic interaction. Thus, possible pairings between the four bases and the 20 residues can be listed ('the chemical code' in Figure 5a; see also Suzuki, 1994). Some residues bind exclusively to one or two of the four nucleotide bases (the specific partners). For example, Ala has a methyl group and can interact only with the T base strongly, the single base having a methyl group in the major groove. But residues such as Ser can bind to any DNA base (the non-specific partners).

The DNA-binding specificity of the examined HTH proteins can be explained very well (Figure 4) by using the same base–residue contacting pattern deduced from the crystal/NMR structures (Figure 5b).

Local DNA-binding geometries of HTH recognition helices are the same. The same arrangement of the three functional positions in the HTH recognition helices and the common base–residue contacting pattern suggest that the same DNA-binding geometry is adopted by the HTH recognition helices. Indeed, we have compared the DNA-binding geometries of HTH recognition helices using their crystal and NMR coordinates (M.Suzuki and M.Gerstein, submitted) and found that HTH recognition helices bind to the DNA with essentially the same local geometry. These helices binding to DNA is not parallel to the DNA major groove but approximately perpendicular to the DNA helix axis (Figure 6a).

Combination of the two recognition helices, which imposes a particular DNA superstructure

Spacing differences make DNA–protein interactions different as a whole. Most HTH proteins form dimers and thus pairs of recognition helices bind to the same DNA. Since we have identified the four base pairs contacted by individual HTH recognition helices (i.e. the 'direct binding sites' shown in Figure 4), the differences in combining pairs of direct binding

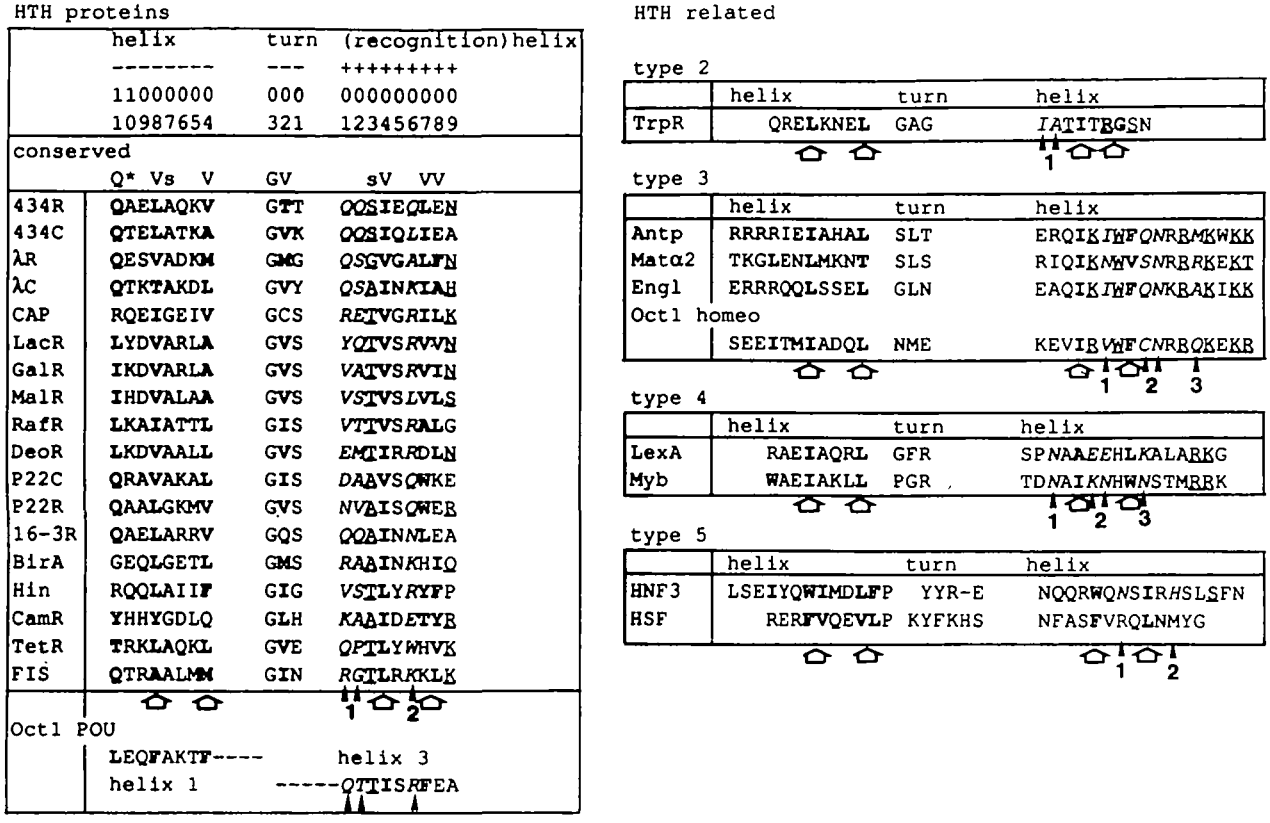


Fig. 3. Sequences of the recognition helices in HTH (left) and related (right) proteins. The residues in a recognition helix can be divided into three types: (i) (facing bases), (ii) (contacting phosphates) and (iii) (facing away from the DNA and packing into the protein). Known or predicted type (i) residues are shown in italic, while known or predicted type (ii) residues in the recognition helices are underlined. Type (iii) residues, which are shown in bold, are usually hydrophobic though they can also be Gln, which contacts another Gln (Pabo *et al.*, 1990). Certain key conserved positions are indicated: those which are consistently occupied by hydrophobic residues (V), those by hydrophobic or Gln (Q*), those by Gly (G) and those by small residues (S), i.e. Gly, Ala, Ser, Cys or Thr. In the proteins listed here, parts of two turns in the recognition helices, which are on the same phase and parts of two turns in the preceding helix, again on the same phase, are occupied by hydrophobic residues. These are marked with open arrows. Up to three turns (marked with 1, 2 or 3) are used for base recognition.

sites into the same DNA can be described (Figure 5c) in terms of two parameters.

(i) The number of base pairs inserted between the two direct binding sites, i.e. the number of 'spacer' base pairs, n, (while the 'spacing' number, N, is defined as the number of base pairs between the centres of the two direct binding sites; N is equal to $n + 4$).

(ii) The direction of the two binding sites relative to the centre of the spacer; this is the same as the N–C direction of the recognition helices relative to the centre of the DNA; if the C-terminus is closer to the centre, the direction is referred to as S^C, otherwise as S^N.

The ways the direct binding sites are arranged in the same DNA and the ways the two recognition helices are combined in the dimers vary between the different HTH proteins. This is the main reason why the overall DNA–protein interactions are different, while the local binding geometries are similar. *Ebg*R and *Gal*R are predicted to recognize the same four base pairs in the direct binding sites, TTAC, but with different spacings (n = 3 for *Ebg*R and n = 4 for *Gal*R). For such discrimination, the spacing differences in the DNAs and the combination differences in the dimers are essential.

Combination parameters can be related to the DNA superstructure. DNA is not necessarily straight but can be curved. Therefore, even with the same number of spacer base pairs different superstructures can be constructed. CAP, 434C and 434R bind to DNA of exactly the same spacing type (i.e. S^C, n = 6), but the DNA which is binding CAP is sharply bent, while those binding 434C and 434R are straighter (Figure 6b).

If the C_α atoms in one of the two recognition helices in CAP are overlapped exactly onto those in one of the helices in 434C, one will notice that the DNA-binding geometry of the overlapped helices relative to the local DNA is very similar (boxed in Figure 6b). However, the non-overlapped helices in the two proteins are positioned rather differently.

The degree to which positioning of the two recognition helices can change a DNA superstructure is limited, e.g. the distance between the two recognition helices cannot be much larger than 3.4 Å×N, but within the limitation the DNA superstructure can be imposed by the positioning of the two recognition helices.

We have characterized the combination of pairs of recognition helices in the crystal/NMR structures (Figure 5d) with the four parameters, l, δ, ε and ω (described in Materials and methods and Figure 2). (Note that some HTH proteins such as 434C dimerize only in the presence of DNA so the parameters cannot be calculated for these proteins in the absence of DNA.) Since the local DNA-binding geometry of a HTH recognition helix is kept the same, these parameters can be directly related to the characteristics of the DNA superstructure (Figure 2).

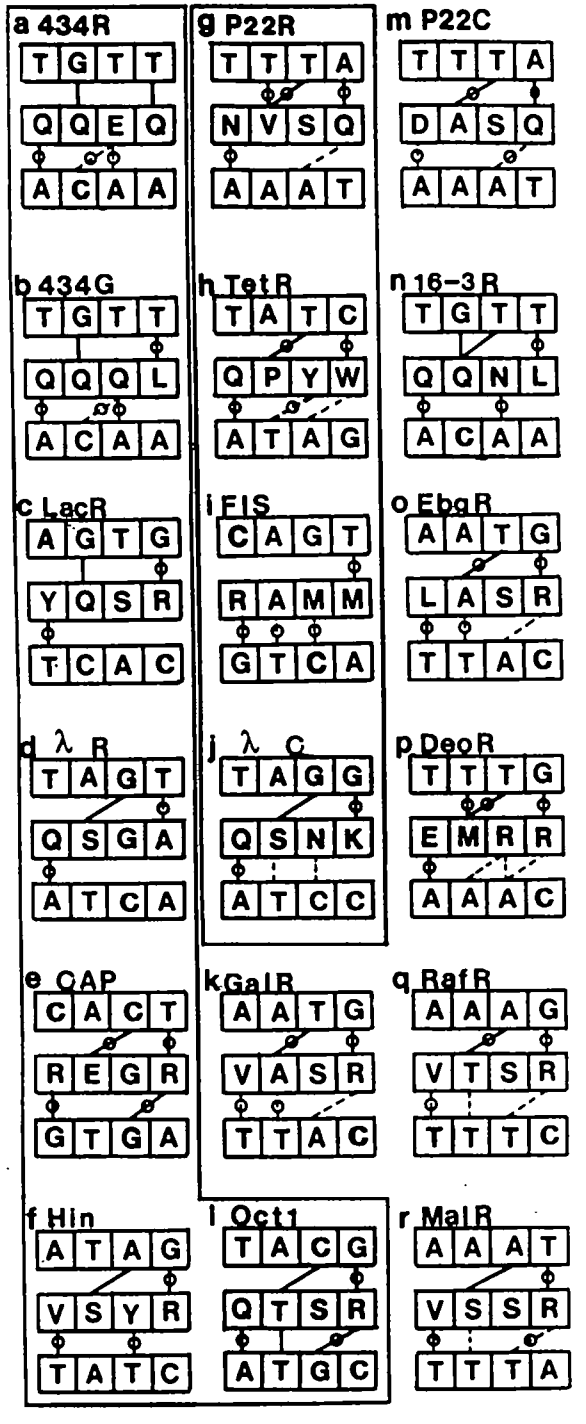
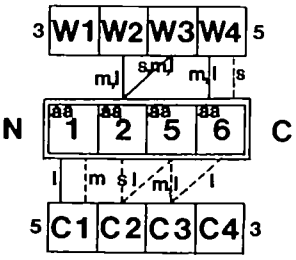


Fig. 4. Contacts between nucleotide bases and amino acid side chains. The figure is drawn in the same way as Figure 5(b). (a)–(f) and (l) Contacts found in crystal and NMR structures. (g)–(j) Contacts predicted for HTH proteins, for which only a structure without DNA is currently known. The structures of the other proteins have not yet been determined. The binding sequences of these proteins on DNA are known (taken from Lehming *et al.*, 1991; Wissmann *et al.*, 1991; Saenger *et al.*, 1993), although the core four base pairs and the N to C direction of the helices relative to DNA are predicted here according to the recognition rules. Gln(aa1) of the Deo repressor cannot bind to T(C1) by a hydrogen bond (j). However, its stem may contact the methyl of T by a hydrophobic interaction. In (a)–(f) and (l) the contacts found in the crystal and NMR structures are shown with solid lines. The dashed lines show contacts that are just beyond the usual thresholds for hydrogen bonds or hydrophobic interaction. In the others the solid and dashed lines are used as the same way as in Figure 5(b). Those which are 'specific' (see Figure 5a) are marked with a circle.

a

	small	medium	large	aromatic
A	Cys,Ser,Thr	Asn,Asp,His	Gln,Glu,Arg,Lys,Met	Tyr,Trp
T	Ala,Cys,Ser,Thr	Val,Ile,Asn,His	Leu,Met,Gln,Arg,Lys	Tyr,Phe,Trp
G	Cys,Ser,Thr	His,Asn	Arg,Lys,Gln	Tyr
C	Val,Cys,Ser,Thr	Asp,Asn,His,Ile	Glu,Gln,Leu,Met	Tyr,Phe,Trp

b



c

Direction	Spacer(n)	Spacing(N)	Protein
S^I	2	6	RafR,TetR
S^I	3	7	EbgR,MalR
S^I	4	8	LacR,GalR
S^C	5	9	FIS
S^I	6	10	DeoR,16-3R
S^C	6	10	CAP,434C,434R
S^C	7	11	λR,λC
S^I	8	12	CytR
S^I	10	14	P22C,P22R

d

	PDB	DNA	d(Å)	ω(°)	ε(°)	δ(°)	n	N
(λR	1LRP	-	35	+17	-42	+62	--	---)
λR	**1LMB**	+	**34**	+22	-7	+11	+7	+11
434R	**2OR1**	+	**28**	+12	-9	+3	+6	+10
434R	**1PER**	+	**28**	+13	-8	+6	+6	+10
434R	**1RPE**	+	**28**	+13	-10	+7	+6	+10
434C	**3CRO**	+	**28**	+13	-6	+7	+6	+10
CAP	3GAP	-	29	-20	+11	+5	---	---
CAP	**1CGP**	+	**31**	-8	+23	+15	+6	+10
FIS	1FIA	-	24	+29	-6	+6	---	---
FIS	3FIS	-	24	+29	-5	+1	---	---

Fig. 5. (a)–(c) DNA recognition rules and (d) combination parameters of HTH proteins. (a) The size and DNA base-binding specificity of amino acid residues can be summarized in the chemical code table. Specific partners (see text) are shown in bold. (b) Contacts between DNA bases and the recognition-helix residues can be summarized into a stereochemical chart. This is essentially a schematic sketch of the part of the DNA major groove to which the recognition helix binds. W and C, the Watson(W) and the Crick(C) strands. s, m and l, the size of residues used for the contacts, small (s), medium (m) and large (l). Solid lines show the contacts frequently observed in the crystal and NMR structures. Dashed lines show contacts observed rarely or which have distances that are slightly beyond conventional definitions of hydrogen bonding or hydrophobic interaction. These latter contacts still might be important for the binding specificity. For example, it has been strongly argued that Ala(aa6)–T(W4) contact is important for DNA discrimination by λR but the distance in the crystal structure is slightly beyond the normal threshold for hydrophobic interaction (see Ptashne, 1992). (c) The spacing between binding sites in HTH proteins. The number of DNA bases between the two direct binding sites (spacer *n* and spacing *N*) and the 5′ to 3′ direction of direct binding site relative to the 2-fold symmetry axis (direction) are shown. The 5′ to 3′ direction of CAP and 434R is referred as to 'S^N', while the opposite is 'S^C'. The spacer base number (*n*) is defined as that between the two direct binding sites, each of which is composed of four base pairs, while the spacing number (*N*) is defined as that between the centres of the two direct binding sites. Consequently, $N = n + 4$. (d) Values of the combination parameters calculated from the crystal structures. The spacer (*n*) and spacing (*N*) numbers are also shown. The values for ε and δ are averaged over the two helices in each HTH protein.

Parameter *l* can be related to the '*N*' spacing between the two binding sites. *l* must be shorter than 3.4 Å×*N*. When $N = 11$, *l* is 34 Å (λR) and when $N = 10$, *l* is 28–31 Å (434R, 434C and CAP).

Parameter δ can be related to the concave/convex curvature of the DNA towards the protein. If δ is positive, the DNA is convex and if it is negative, the DNA is concave (Figure 2d). In other words, when δ is positive the protein is placed inside the DNA superhelix, whereas when it is negative, the protein is placed outside.

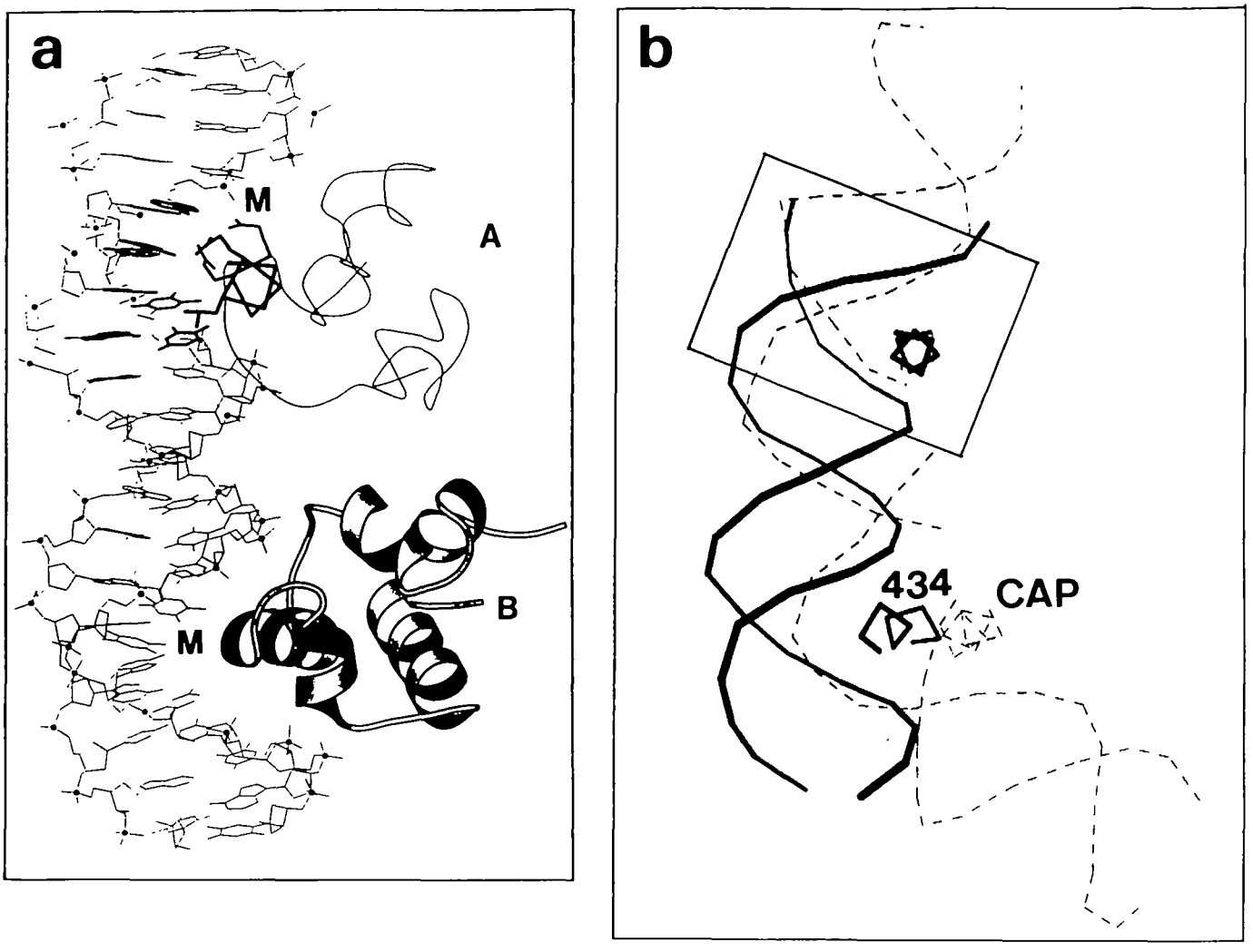


Fig. 6. DNA-binding by HTH proteins. (**a**) DNA-binding by 434C (PDB code, 3CRO). The four base pairs contacted by one of the two monomers (A) is highlighted in bold. Note that the recognition helices run perpendicular to the axis of the DNA double helix. The DNA major groove is marked M. (**b**) Different ways of combining two recognition helices in 434cro (3CRO, shown with solid lines) and CAP (1CGP, shown with dashed line). Note that although the overall bindings of CAP and 434C are very different, the local binding geometry of the overlapped helices is very similar (indicated by the box).

Parameter ω can be related to the helical twist of the DNA (Figure 2b and c). If ω is positive, the DNA is more twisted between the two binding sites and if it is negative, it is less twisted or untwisted.

Parameter ε can be related to the handedness of the superhelicity of the DNA (Figure 2d–f). If ε and δ have the same sign, the DNA superhelix is left-handed (Figure 2e), while if these have opposite signs, the DNA superhelix is right-handed.

CAP and FIS induce different DNA superstructures. The combination parameters calculated for CAP are distinctively different from those of the other structures (Figure 5d). In particular, ε is positive and large, while the other structures have negative ε. Together with the positive and large value for δ, this value of ε implies that CAP induces a left-handed superhelical structure onto DNA, placing the protein inside the superhelix (Figure 2e). Parameter ω is negative in CAP and positive in the other HTH structures. This implies that CAP untwists the DNA and, indeed, the average helical twist between the centres of the two direct binding sites in the CAP structure is found to be 30° per base pair, while in the other three structures (DNA-λR, DNA-434C and DNA-434R) it is 34–35° (Figure 7).

FIS has been crystallized only in the absence of DNA but DNA–FIS complexes have been modelled (Kostrewa *et al.*, 1991; Yuan *et al.*, 1991). By examining changes in the parameters of CAP and λR upon binding to DNA, the degree to which the parameters of FIS change upon DNA-binding may be estimated to be small. Distance *l* of FIS, 24 Å, is smallest among the HTH structures. Thus, the spacing (characterized by numbers *n* and *N*) is expected to be smaller for FIS than for the other structures (i.e. $n = 5$, $N = 9$ for FIS). This prediction coincides with the models (Saenger *et al.*, 1993).

Parameter ω of FIS is the largest among those calculated. This implies that when FIS binds to DNA, it probably twists the DNA more. This can be significant, because the spacing of FIS (*n* and *N*) is predicted to be even smaller than those of the others. Thus, the DNA superstructure imposed by FIS is expected to be different from that induced by CAP.

Two TG/CA steps which adopt a distinctive conformation upon binding CAP

For an understanding of the structural features of DNA which are important for inducing a particular superstructure to follow a protein surface, we have calculated the six parameters at

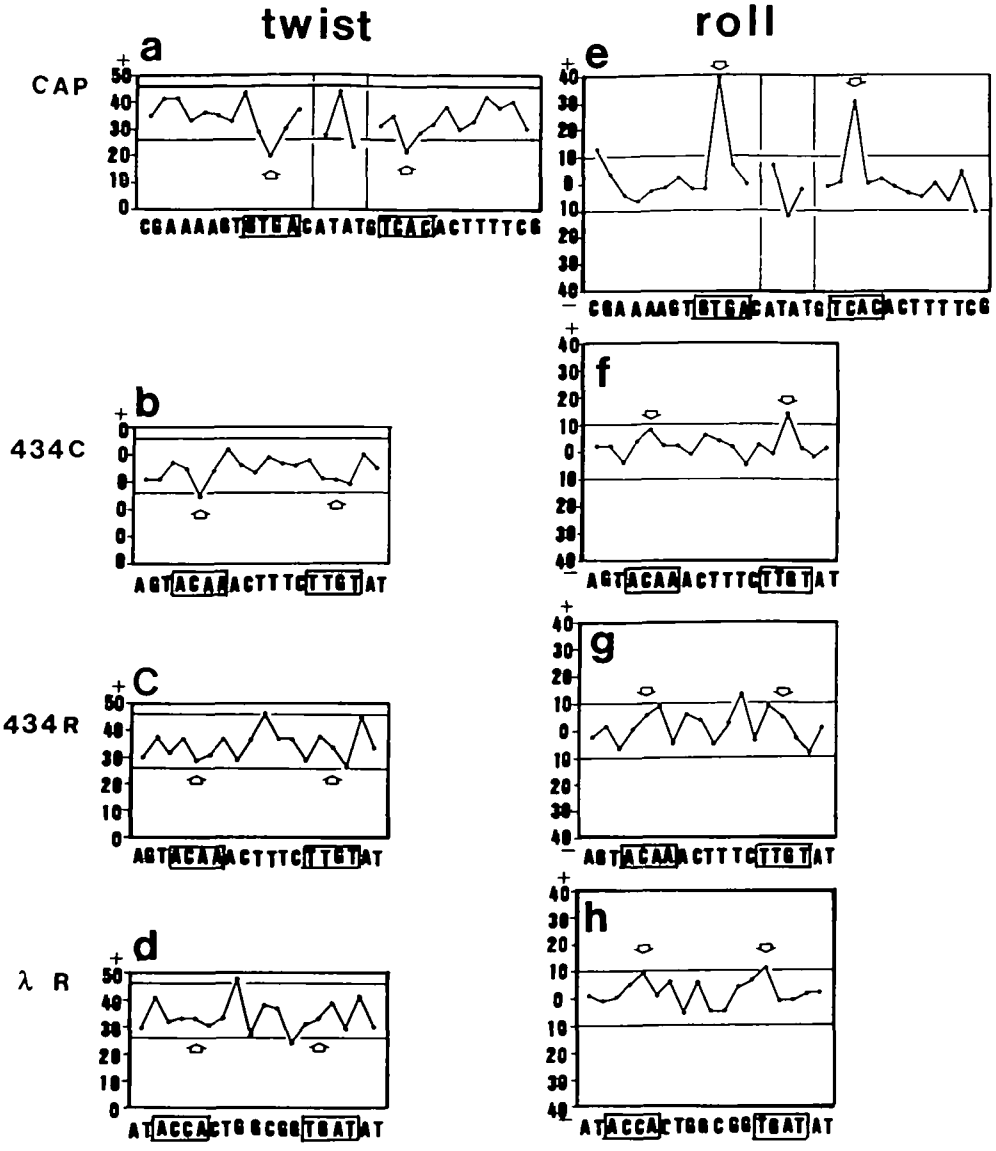


Fig. 7. (a)–(d) Helical twist and (e)–(h) roll of each base pair step in CAP (a and e), 434C (b and f), 434R (c and g) and λR (g and h). The parameters were calculated using crystal structures of DNA-CAP (1CGP), DNA-434R (2OR1), DNA-434C (3CRO) and DNA-λR (1LMB). The DNA in the CAP structure has nicks at two positions, consequently, the plots have discontinuities marked by vertical lines. In each direct binding site, the four consecutive bases interacting with the recognition helix are shown boxed and the the TG/CA sequences are marked with arrows.

each base pair step: helical twist, roll, tilt, shift, slide and rise in the complex with CAP (1CGP), 434R (2OR1), 434C (3CRO) and λR (1LMB) (Figure 7).

We found that the DNA bending by CAP is achieved by untwisting and rolling of the TG/CA steps at the centres in the two direct-binding sites (Figure 7a and e). The rolling, which opens the steps on the major groove side, bends the DNA by ~40° (Schultz *et al.*, 1991) at each step, while untwisting introduces left handedness into the overall DNA superhelix (Figure 8b).

The POU domain of Oct-1 binds to a sequence which has the same TG/CA step at the equivalent position. Although the full details have not been published, Oct-1 POU bends the DNA, again by ~40° (Klemm *et al.*, 1994). Other crystal structures of HTH proteins (434C, 434R and λR) have TG/CA steps in their direct binding sites but not at the equivalent position with the same orientation as in the CAP structure. These DNA structures are not bent as sharply as the CAP structure. However, although the changes are much smaller and insignificant, the TG/CA steps in 434C, 434R and λR do have a tendency to untwist and roll (marked with arrows in Figure 7).

As shown in Figure 8(a), the untwisting and the rolling can be explained as part of one overall movement in the TG/CA steps. Two important characteristics of a TG/CA step can explain why it, rather than other sequences, particularly facilitates this untwisting and rolling.

First, a HTH recognition helix is closest to base pairs 2 and 3 in the binding site (Figure 6a). The recognition helix axis runs approximately perpendicular to the DNA helix axis and approximately parallel to the edges of base pairs 2 and 3 on the major groove side. The shape of two consecutive base pairs on the major groove side is dependent on the base sequence (Figure 9a and b). The pyrimidine bases (T and C) are bulkier towards the major groove than the purine bases (A and G). The T base is especially bulky (Figure 9c and d). As a consequence, the edges of the two base pairs in pyrimidine–purine steps (TG/CA, TA, CG) open widely on the major

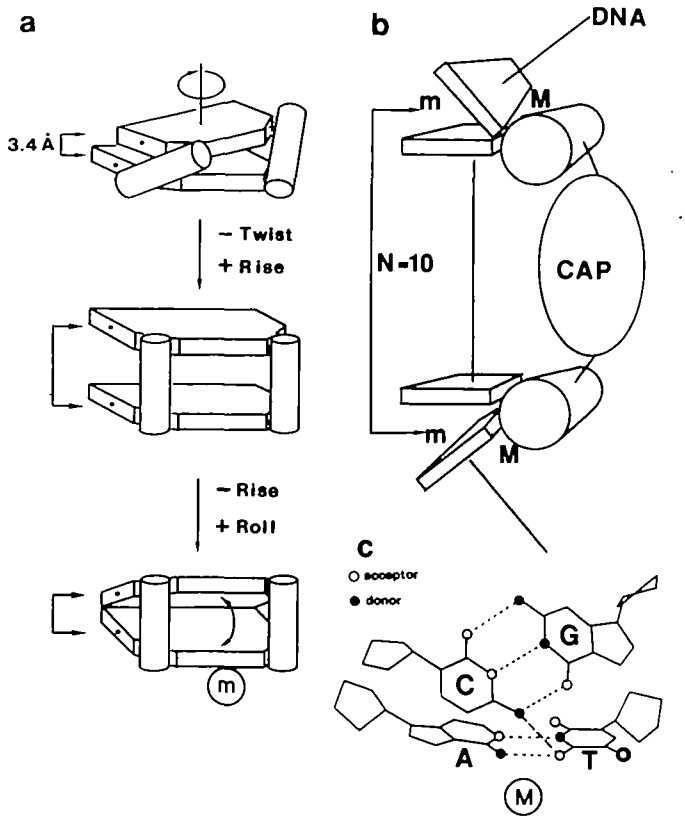


Fig. 8. Untwisting–rolling of the TG/CA step upon binding CAP. (**a**) How a base step can be untwisted and rolled. The sugar–phosphate backbones are closer to the DNA minor groove. The backbone lengths are larger than the stacking distance of the two base pairs, which makes DNA helically twisted (top). When the two base pairs are untwisted the rise increases (middle). However, a high rise is not appropriate for base stacking. One way of improving the stacking is to move the bases closer on the major groove side by rolling the base step (bottom). The DNA base pairs are looked at from the minor (m) groove side. (**b**) How the untwisting and the rolling of the two TG/CA steps in the direct binding sites are combined to bend the DNA by CAP. Compare this figure with Figures 2(e) and 6(b). (**c**) The interbase pair hydrogen bond that facilitates the untwisting and rolling of TG/CA step in the complex with CAP (1CGP). The hydrogen bond is made between the N4 of C and the O4 of T. The step is looked at from the major (M) groove side.

groove side (Figure 9a). The untwisting–rolling appears to be a consequence of the TG/CA step for better following the recognition helix. That is, the untwisting aligns the edges of the two base pairs closer and more parallel to each other so that these can accept the straight α-helix (Figure 8b) and this facilitates the rolling (Figure 8a).

Second, the untwisting–rolling of the step seems to be stabilized by an inter-basepair hydrogen bond on the major groove side between T and C (Figure 8c; although this hydrogen bond was not mentioned in the original paper, we have identified it using the crystal coordinates). Such an inter-basepair hydrogen bond can be made only when T and G, which have hydrogen bond acceptors on the major groove side, are placed on the same DNA strand and A and C, which have hydrogen bond donors, are placed on the other DNA strand. TG/CA is one of the four such possible combinations: TG/CA, GT/AC, TT/AA and GG/CC. Thus, a TG/CA step is the only step which has both features; it is a pyrimidine–purine step and T and G are placed on the same side.

HTH fold and the recognition helix

A recognition helix cannot be totally independent from the protein fold to which it is incorporated, but the protein fold and the type of a recognition helix do not have a simple one-to-one correspondence. The HTH protein fold can adopt at least five different types of recognition helices (Figure 3). Obviously the first type has been decribed in detail above. This type can be incoporated into another protein fold, that of the POU domain (Figure 4l; note that the contacting pattern of POU is the same as that of other HTH proteins of this type).

The second type of recognition helix is that of TrpR. This type cannot adopt the same binding geometry as the first type, since positions 5 and 6 are part of a pocket holding the co-factor, tryptophan, (these positions are regarded as being of type iii) and, thus, cannot face the DNA bases as in the first type (which are regarded as being of type ii).

The third type of recognition helix is the probe helix type found in homeodomains. In this type the positions used for contacting the DNA bases are shifted to the C-terminus compared with those in a classic HTH recognition helix (Figure 3). It has some basic residues which contact DNA phosphates at the C-terminus. As a consequence, a homeo helix adopts a different DNA-binding geometry (Suzuki, 1993, 1994; M.Suzuki and M.Gerstein, submitted).

The fourth type of recognition helix is found in the transcription factors Myb and LexA. The structure of the third DNA-binding domain of Myb (Ogata *et al.*, 1992) and the DNA-binding domain of LexA (Fogh *et al.*, 1994) have been determined by NMR and these have the HTH fold. The sequences, structures and DNA-binding modes predicted from biochemical experiments of the two proteins are very similar (Suzuki, 1995).

The fifth type of recognition helix seems to be that found in HNF3 and heat shock factor (HSF). These proteins can be grouped into the 'winged helix' family (Clark *et al.*, 1993). Although the term, the winged helix family, was originally used for a larger group which included CAP and the homeodomain, here we use it for a smaller group.

Conclusion

In this paper we have discussed three major aspects of DNA–HTH protein interactions: (i) that individual recognition helices of classic HTH proteins bind to DNA in the same way, (ii) that pairs of HTH recognition helices can be combined in different ways in dimers and (iii) that these differences are used for discriminating between DNAs which have different superstructures, in particular, different numbers of base pairs between the two direct binding sites.

The DNA-binding mode of a recognition helix is more loosely related to the overall protein fold than was once expected. Consequently, when the term HTH is used, it should be specified whether it is being used to denote the particular protein fold or the particular type of recognition helix (in this paper it usually means a protein which has the HTH fold and a recognition helix of the first type). Such specification will clear up much of the confusion created by the complication.

Acknowledgements

We thank Professor W.Olsen for giving us her computer program. We thank Mr S.E.Brenner for his help in preparing the figures. We thank Drs J.Finch and C.Chothia for their critical reading of the manuscript. M.G. acknowledges support from a Damon Runyon–Walter Winchell fellowship (DRG-1272).

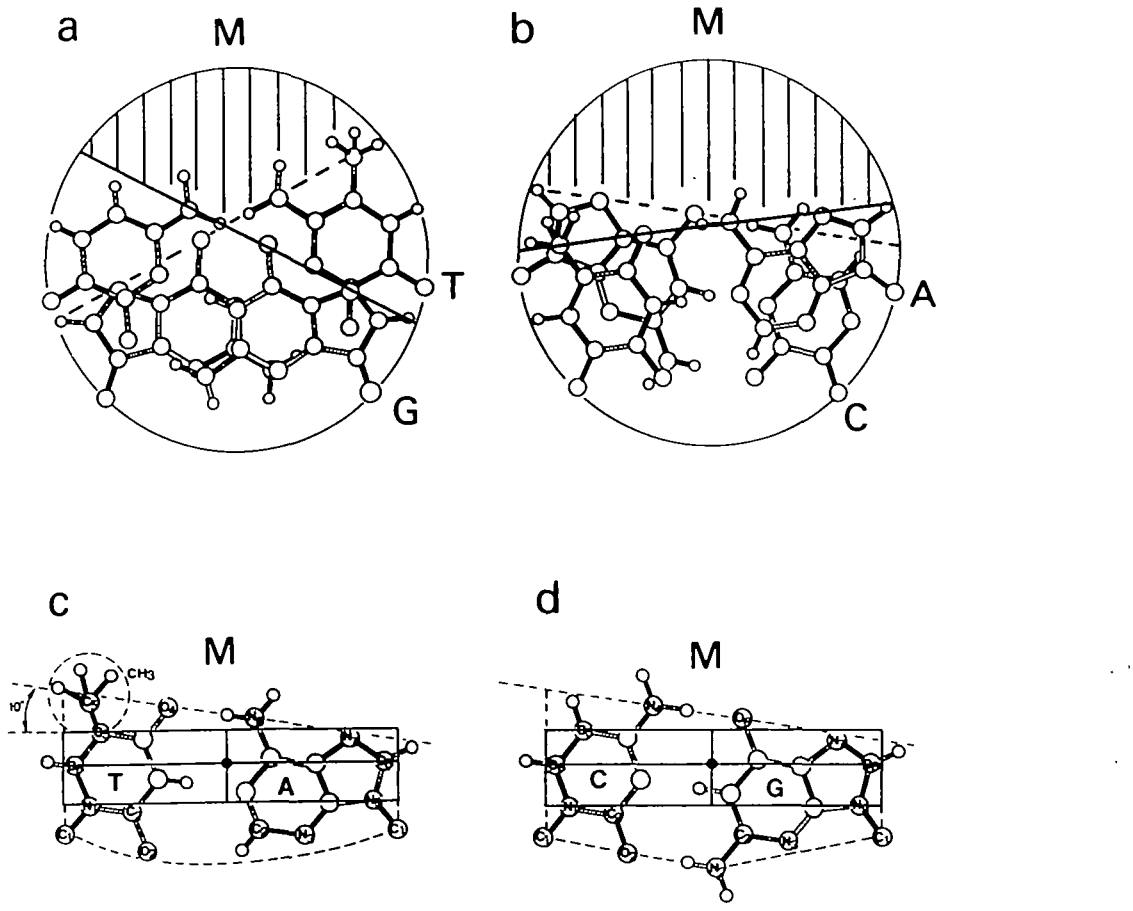


Fig. 9. (a) and (b) Features of DNA used for bending by CAP. The size of the opening of two consecutive base pairs towards the major groove (M) is dependent on positioning of the purine and pyrimidine bases. If the helical twist is kept the same, a pyrimidine–purine step (TG/CA in a) creates the major groove narrower than that made by a purine–pyrimidine step (AC/GT in b). The edges of the two base pairs in TG/CA are rotated from each other by ~60°, while those in AC/GT by ~20°. (c) and (d) The differences in opening in (a) and (b) can be understood by the fact that the pyrimidine residues, especially T, project bulky groups towards the major groove. These bulky projections are particularly evident if one draws tilted lines connecting the chemical features on the major groove side of the DNA bases

References

Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) *Nature*, **242**, 899–907.

Anderson,W.F., Ohlendorf,D.H., Takeda,Y. and Matthews,B.W. (1981) *Nature*, **290**, 754–758.

Anderson,J.E., Ptashne,M. and Harrison,S.C. (1987) *Nature*, **326**, 846–852.

Assa-Munt,N., Mortishire-Smith,R.J., Aurora,R., Herr,W. and Wright,P.E. (1993) *Cell*, **73**, 193–205.

Babcock,N.S., Pednault,E.P.D. and Olson,W.K. (1993) *J. Biomol. Struct. Dynam.*, **11**, 597–628.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Billeter,M., Qian,Y.Q., Otting,G., Müller,M., Gehring,W. and Wüthrich,K. (1993) *J. Mol. Biol.*, **234**, 1084–1094.

Branden,C. and Tooze,J. (1991) *Introduction to Protein Structure*. Garland Publishing Inc., New York.

Brennan,R.G. (1991) *Curr. Opin. Struct. Biol.*, **1**, 80–88.

Brennan,R.G., Roderick,S.L., Takeda,Y. and Matthews,B.W. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 8165–8169.

Chuprina,V.P., Rullmann,J.A.C., Lamerichs,R.M.J.N., Van Boom,J.H., Boelens,R. and Kaptein,R. (1993) *J. Mol. Biol.*, **234**, 446–462.

Clark,K.L., Halay,E.D., Lai,E. and Burley,S.K. (1993) *Nature*, **364**, 412–420.

Clarke,N.D., Beamer,J.L., Goldberg,H.R., Berkower,C. and Pabo,C.O. (1991) *Nature*, **254**, 267–270.

Dekker,N., Cox,M., Boelens,R., Verrijzer,C.P., van der Vliet,P.C. and Kaptein,R. (1993) *Nature*, **362**, 852–855.

Diekmann,S. (1989) *EMBO J.*, **8**, 1–4.

Feng,J.-A., Johnson,R.-C. and Dickerson,R.E. (1994) *Science*, **263**, 348–355.

Fogh,R.H., Ottleben,G., Rüterjans,H., Schnarr,M., Boelens,R. and Kaptein,R. (1994) *EMBO J.*, **13**, 3936–3944.

Harrison,C.J., Bohm,A.A. and Nelson,C.W. (1994) *Science*, **263**, 224–227.

Hinrichs,W., Kisker,C., Düvel,M., Müller,A., Tovar,K., Hillen,W. and Saenger,W. (1994) *Science*, **264**, 418–420.

Jordan,S.R. and Pabo,C.O. (1988) *Science*, **242**, 893–899.

Kissinger,G.R., Liu,B., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) *Cell*, **63**, 579–590.

Kisters-Woike,B., Lehming,N., Sartorius,J., von Wilcken-Bergmann,B. and Müller-Hill,B. (1991) *Eur. J. Biochem.*, **198**, 411–419.

Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,C.O. (1994) *Cell*, **77**, 21–32.

Kostrewa,D.K., Granzin,J.G., Koch,C., Choe,H.-W., Raghunathan,S., Wolf,W., Labahn,J., Kahmann,R. and Saenger,W. (1991) *Nature*, **349**, 178–180.

Lamerichs,R.M.J.N., Padilla,A., Boelens,R., Kaptein,R., Ottleben,G., Rüterjans,H., Granger-Schnarr,M., Oertel,P. and Schnarr,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 6863–6867.

Lawson,C.L. and Carey,J. (1993) *Nature*, **366**, 178–182.

Lehming,N., Sartorius,J., Kisters-Woike,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1991) In Ecstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology 5*. Springer-Verlag, Berlin, pp.114–125.

Matthews,B.W. (1988) *Nature*, **335**, 294–295.

Mondragón,A. and Harrison,S.C. (1991) *J. Mol. Biol.*, **219**, 321–334.

Mondragón,A., Wolberger,C. and Harrison,S.C. (1989a) *J. Mol. Biol.*, **205**, 179–188.

Mondragón,A., Subbiah,S., Almo,S.C., Drottar,M. and Harrison,S.C. (1989b) *J. Mol. Biol.*, **205**, 189–200.

Neri,D., Billeter,M. and Wüthrich,K. (1992) *J. Mol. Biol.*, **223**, 743–767.

Ogata,K., Hojo,H., Aimoto,S., Nakai,T., Nakamura,H., Sarai,A., Ishi,S. and Nishimura,Y. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 6428–6432.
Ohlendorf,D.H., Anderson,W.F., Fisher,R.G., Takeda,Y. and Matthews,B.W. (1982) *Nature*, **298**, 718–723.
Otwinowski,Z., Schevitz,R.W., Zhang,R.-G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) *Nature*, **335**, 321–329.
Pabo,C.O. and Lewis,M. (1982) *Nature*, **298**, 443–447.
Pabo,C.O. and Sauer,R.T. (1984) *Annu. Rev. Biochem.*, **53**, 293–321.
Pabo,C.O., Aggarwal,A.K., Jordan,S.R., Beamer,L.J., Obeysekare,U.R. and Harrison,S.C. (1990) *Science*, **247**, 1210–1213.
Ptashne,M. (1992) *A Genetic Switch*. 2nd edition. Cell Press, Cambridge, MA.
Rodgers,D.W. and Harrison,S.C. (1993) *Structure*, **1**, 227–239.
Saenger,W., Sandmann,C., Theis,K., Starikov,E.B., Kostrewa,D., Labahn,J. and Granzin,J. (1993) In Ecstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology 7*. Springer-Verlag, Berlin, pp. 158–169.
Sauer,R.T., Yocum,R.R., Duulittle,R.F., Lewis,M. and Pabo,C.O. (1982) *Nature*, **298**, 447–451.
Schevitz,R.W., Otwinowski,O., Joachimiak,A., Lawson,C.L. and Sigler,P.B. (1985) *Nature*, **317**, 782–786.
Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) *Nature*, **253**, 1001–1007.
Sevilla-sierra,P., Otting,G. and Wüthrich,K. (1994) *J. Mol. Biol.*, **235**, 1001–1021.
Shimon,L.J.W. and Harrison,S.C. (1993) *J. Mol. Biol.*, **232**, 826–838.
Steitz,T.A. (1993) *Structural Studies of Protein–Nucleic Acid Interaction*. Cambridge University Press, Cambridge.
Suzuki,M. (1993) *EMBO J.*, **12**, 3221–3226.
Suzuki,M. (1994) *Structure*, **2**, 317–326.
Suzuki,M. (1995) *Proc. Jap. Acad.*, **B71**, 217–231.
Suzuki,M. and Yagi,N. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
Weber,I.T. and Steitz,T.A. (1987) *J. Mol. Biol.*, **198**, 311–326.
Wilson,K.P., Shewchuk,L.M., Brennan,R.G., Otsuka,A.J. and Matthews,B.W. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 9257–9261.
Wissmann,A., Baumeister,R., Müller,G., Hecht,B., Helbe,V., Pfeleiderer,K. and Hillen,W. (1991) *EMBO J.*, **10**, 4145–4152.
Wolberger,C., Dong,Y., Ptashne,M. and Harrison,S.C. (1988) *Nature*, **335**, 789–795.
Wolberger,C., Vershon,A.K., Liu,B., Johnson,A.D. and Pabo,C.O. (1991) *Cell*, **67**, 517–528.
Yuan,H.S., Finkel,S.E., Feng,J.-A., Kaczor-Grzeskowiak,M., Johnson,R.C.and Dickerson,R.E. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 9558–9562.
Zhang,R.-G., Joachimiak,A., Lawson,C.L., Schevitz,R.W., Otwinowski,Z. and Sigler,P.B. (1987) *Nature*, **327**, 591–597.

Received November 8, 1994; revised January 12, 1995; accepted January 23, 1995

Note added in proof

The crystal structure of the PurR–DNA complex has been determined recently [Schumacher *et al.* (1994) *Science*, **266**, 763–770]. The residue–DNA base contacts found within this structure fall into the pattern described in this paper.